
評估「資料探索者」：數位時代認證道德情報的策略評估

**BardiVarian Twin Stars 在 PeaceCasts4Good 節目中的評論**

**https://archive.org/details/navigating-the-digital-storm-how-data-spelunkers-combat-ai-driven-disinformation**

**PM Thomas，博士，數據侍酒師https://bit.ly/MyPrivateAi**

執行摘要

到2025年夏天，人工智慧生成內容的自我複製和激增將引發一場前所未有的“數據雪崩”，從根本上挑戰全球資訊生態系統的完整性。本報告評估了「數據探索者」作為應對這一重大威脅的解決方案，並特別關注其提供的「經驗證的倫理情報」。分析表明，「數據探索者」提出了令人信服的價值主張，將深厚的數據取證和真實性驗證技術能力與對倫理人工智慧原則的堅定承諾相結合。他們與全球和平與安全的總體目標的戰略契合，使其成為情報機構的積極主動的資產。

儘管在打擊人工智慧驅動的虛假資訊以及在政府內部整合人工智慧解決方案方面存在著重大的技術和組織障礙，「數據探索者」強調可驗證的方法、多元化的專業知識和道德差異化，這使得他們成為值得信賴且至關重要的合作夥伴。主要發現強調了專業知識在應對複雜數位環境方面的必要性、合乎道德的方法的戰略優勢，以及先進的技術解決方案與客

戶機構內部強大的組織準備之間的相互依存關係。建議強調深化試點計畫、擴大合乎道德的人工智慧領導力，並協助機構應對內部實施挑戰，以最大限度地發揮其影響力。

## 1. 引言：人工智慧驅動的假資訊模式和“數據雪崩”

全球資訊環境正處於關鍵時刻，預計 2025 年夏季將出現前所未有的由自我複製的人工智慧 (AI) 產生的內容「資料雪崩」。[1]這場即將到來的資訊洪流引發了人們對資料真實性以及虛假資訊廣泛快速傳播的深切擔憂。這些合成資訊的龐大數量和極快傳播速度預計將壓倒傳統的情報收集方法，形成所謂的「數位戰爭迷霧」或「合成資訊海嘯」。[1]在這樣的環境下，辨別真假變得極為困難，需要專門的數位資訊驗證知識。[1]

假資訊問題一直是人類社會面臨的古老挑戰，而人工智慧的最新進展又使這個問題變得更加嚴重。[2]現代人工智慧工具現在可以毫不費力地創建與真實內容越來越難以區分的虛假圖像和新聞。[2]其中包括極具說服力的深度偽造視頻，描繪公眾人物從事虛假行為，以及人工智能生成的文章，以驚人的準確度精心模仿可靠的新聞來源。[3]報告顯示，到 2023 年，人工智慧支援的假新聞網站數量將增加十倍，其中

許多網站在極少的人工監督下運行，這突顯了這種威脅的規模。[2]一個人的臉或聲音可以如此輕鬆地被複製和編輯，這標誌著現實與虛擬之間的界限正在迅速消失的新時代的到來。[4]

人工智慧產生的假新聞和虛假資訊的氾濫已對關鍵領域造成重大影響，其影響範圍已超越公眾認知，甚至延伸至切實的經濟和政治不穩定。例如，一則關於美國證券交易委員會（SEC）批准比特幣交易所交易基金（ETF）的虛假報道，導致比特幣價格大幅波動。[4]同樣，一張由人工智慧生成的虛構圖像顯示五角大廈附近的一棟建築物被黑色火焰吞沒，導緻美國股市動盪。[4]除了這些備受矚目的事件之外，虛假資訊還損害了醫療機構和專業組織的信譽，網路上一些未經證實的說法宣傳未經證實甚至非法的治療方法。[4]

這種廣泛傳播的假訊息從根本上削弱了公眾對媒體、政府機構和民主話語基礎的信任。[5]人工智慧驅

動的虛假資訊的速度、規模和複雜性，綜合起來，在資訊生態系統中造成了系統性脆弱性。海量的合成資訊足以淹沒並有效破壞公共資訊環境，其速度遠超其被消除的速度。這不僅代表個別假訊息的挑戰，也代表著對資訊本身完整性的根本威脅，使其從媒體素養問題上升為國家安全當務之急。

## 2.「資料洞穴探險者」的價值主張：核心理念與使命

為了應對不斷升級的“數據雪崩”，“數據探索者”將自己定位為關鍵的對策，並提供了一種稱為“經過驗證的道德情報”的獨特解決方案。[1]這項核心服務被定義為經過驗證、值得信賴且符合道德規範的訊息，對於維護全球和平與安全至關重要。[1]該組織自稱是一支專家團隊，擁有專業技能和工具，旨在「深入」混沌的數位環境，提取真知灼見，並嚴格驗證其來源和完整性。他們的專長在於駕馭複雜、多層次

的人工智慧生成內容環境，從無所不在的數位噪音中識別出真正的訊號。[1]

「資料探索者」價值主張的一個顯著特徵是明確地將「和平」作為其驅動使命。這一目標已融入其「經認證的、符合道德的和平情報機構」框架的根本框架之中，使其意圖從一開始就清晰明確。[1]他們的服務與授權情報機構做出明智的決策直接相關，這些決策可以積極防止衝突、緩和緊張局勢並保護弱勢群體，從而直接促進全球穩定。[1]透過調整語言以符合情報機構既定的優先事項（包括國家安全、預防衝突、反恐和維護全球穩定），這種聯繫得到了進一步加強。[1]透過與這些高層目標保持一致，「數據探索者」將他們的專業服務與更廣泛的和平追求隱性地聯繫起來。

「數據探索者」所採用的策略訊息始終強化著這項使命。其提出的口號和關鍵訊息點明確地體現了「和平」的目標，例如「為全球和平提供真實情報：數據探

索者的承諾」和「我們為合乎道德的情報機構提供可驗證的數據，以促進和平與安全」。[1]這種始終如一的溝通確保了和平目標深植於其核心敘事之中。「數據探索者」的核心使命圍繞著“和平”和“預防衝突”，戰略上將自己定位於不僅僅是被動的威脅檢測服務，更是主動的情報資產，直接貢獻於國家安全和外交政策等更高層次的戰略目標。這種方法將他們的價值主張從純粹的技術解決方案提升為策略穩定的基礎要素，從而吸引最高層。。

**3.** 專業能力與方法

「數據探索者」憑藉一系列專業能力和方法脫穎而出，旨在應對人工智慧驅動的虛假資訊的複雜性。他們的核心服務是深入探索混亂數據環境的專業知識。[1]這種能力涵蓋進階資料取證、異常檢測和模式識別，這些對於在海量數位環境中識別人工智慧生成的內容、深度偽造和其他形式的操縱資訊至關重要。[1]他們的技術人員擅長分析細微的數位痕跡，識別異常的數據行為，並辨別表明內容被人為創建或更改的重複結構或異常。[1]

強大的真實性驗證框架是對這種深度專業知識的補充。「資料探索者」利用專有或專門的方法來驗證資料的來源、作者身份和完整性。[1]這些框架採用了多種先進技術：

- 加密技術：加密、數位簽章或雜湊的應用可確保資料完整性並確認其來源，從而為資訊提供安全保管鏈。[1]
- 數位浮水印分析：這涉及檢測和分析數位內容中嵌入的隱藏資訊的能力，這對於追蹤其來源或驗證其真實性至關重要。[1]
- 資料流的行為分析：透過監控資料流動時的模式和特徵，團隊可以識別暗示操縱或非人為來源的偏差，從而提供合成內容的早期預警。[1]
- 與可靠來源交叉引用：關鍵步驟是將潛在的可疑數據與已知可靠來源的資訊進行比較，以驗證其準確性和完整性，並利用既定的真實基準。[1]

其核心優勢和承諾的體現在於，他們始終堅守人工智慧的道德原則。這項承諾涵蓋兩個關鍵面向：首先，積極識別惡意使用人工智慧的情況，例如將人工

智慧用於製造虛假資訊等有害目的的情況。[1]第二，同樣重要的是確保自身運作中的行為符合道德規範。他們的運作方式透明、公正，並嚴格尊重隱私和人權。[1]「資料探索者」明確地與既定的人工智慧倫理框架保持一致，甚至經常超越這些框架，包括聯合國教科文組織的《人工智慧倫理建議》和情報界的《人工智慧倫理框架》。[1]這種積極主動地負責任地處理和分析數據的方法，確保了其運作符合最高的道德標準。負責任的人工智慧原則，例如公平、透明、問責和安全，被強調為建立公眾信任和製定有效的反虛假資訊措施的關鍵。[5]此外，可解釋人工智慧 (XAI) 的應用對於闡明其檢測模型如何運作至關重要，使利害關係人能夠了解技術流程及其對決策的影響。[5]

下表清晰簡潔地概述了「資料探索者」行動的技術和倫理支柱，快速展現了他們專業知識的深度和廣度。對於高層決策者而言，它展現了他們應對人工智慧驅動的虛假資訊複雜挑戰的能力，同時也凸顯了他們對倫理標準的承諾，這對於維護人們對人工智慧系統的信任至關重要。

表1:「資料洞穴探險者」核心能力及驗證技術

| 能力領域 | 關鍵技術/方法 | 目的/好處 |
| --- | --- | --- |
| 深層專業知識 | 進階資料取證、異常檢測、模式識別 | 識別人工智慧產生的內容、深度偽造和操縱的資訊;分析數位痕跡和異常數據行為。 |
| 真實性驗證框架 | 密碼技術(加密、雜湊、數位簽章)、數位浮水印分析、資料流行為分析、與可信任來源的交叉引用 | 驗證資料來源、作者和完整性;確保不被篡改並確認來源;追蹤來源並驗證準確性。 |
| 人工智慧道德原則 | 識別惡意人工智慧的使用、透明度、公正性、尊重隱私/人權、與人工智慧道德框架保持一致(例如,聯合國教科文組織、IC 人工智慧倫理)、可解釋人工智慧(XAI) | 檢測有害的人工智慧部署;確保負責任的資料處理;透過清晰的決策過程建立信任;防止歧視性結果。 |

「資料探索者」對人工智慧倫理的強調，不僅僅是出於道德或合規方面的考慮；它代表了任何情報領域實體的戰略要務。研究表明，存在偏見或不透明的人工智慧系統可能會有意或無意地壓制合法內容，放大有害的錯誤訊息，並削弱公眾信任。[5]歷史上的案例，例如演算法偏見導致自動評估系統出現歧視性結果，已經引起了公眾的強烈反對和國家醜聞。[6]公眾信任對於情報行動的有效性和民主話語的健康至關重要。[5]因此，「資料探索者」的道德承諾直接影響他們所提供情報的可靠性和可信度，使其成為其價值主張的關鍵組成部分和機構採用的先決條件。

**4.** 在情報界建立信任和信譽

在情報界，建立和維護信任至關重要，因為情報界的風險本身就很高。「資料探索者」提出了一種多管齊下的方法來建立可信度。其中一項關鍵策略是展示經過驗證的方法。雖然具體的工具可能是專有的，但「數據探索者」計劃闡明類型他們採用了一系列先進的方法。例如「對抗性人工智慧檢測」、「可解釋的真相識別人工智慧」（XAI）和「基於區塊鏈的資料

來源追蹤」。[1]這種方法在不洩漏敏感細節的情況下，展現了技術的精湛和創新的立場。可解釋的人工智慧 (XAI) 在此背景下尤其重要，因為它清晰地闡述了檢測和歸因模型的運作方式，使利害關係人能夠理解技術流程及其對決策的影響。[5]強大的文檔和可追溯性也是不可或缺的，可以實現對人工智慧系統的有效審計和治理。[5]

該提案非常重視團隊的多元化專業技能。「資料探索者」將展示其團隊的跨學科技能，其中包括資料科學家、網路安全專家、語言學家、區域專家、倫理學家和情報分析員。[1]這種多樣性體現了他們對虛假資訊問題複雜性和多面性的全面理解，這超越了單純的技術層面，涵蓋了語言、文化和倫理方面的考量。此外，他們強調團隊在「高風險環境」中的經驗，這直接反映了情報機構的行動現實和需求，顯示了他們在壓力下的準備和可靠性。[1]

或許，最有效的即時建立信任的策略是提供試點計畫和演示。「數據探索者」提出了一些可控的試點項目，讓各機構能夠利用自身數據挑戰或模擬場景，親眼見證其能力。[1]這種實際演示提供了切實的價值證明，透過在受控的低風險環境中直接評估有效

性、準確性和道德遵守情況，將最初的懷疑轉化為信心。[1]

為了建立長期信譽，「數據探索者」計畫積極參與思想領導力建設和策略合作。這包括發佈白皮書、積極參與會議，以及參與人工智慧倫理、數據真實性和智慧未來的更廣泛討論。[1]這些活動使他們成為這一新興領域的權威人物和領導者。此外，探索與知名學術機構、符合道德規範的人工智慧智庫以及知名網路安全公司的合作，可以進一步提升他們的信譽，擴大他們的影響力，並利用外部驗證。[1]在人工智慧承諾可能被誇大甚至欺詐的市場中，透過試點計畫獲得實際可驗證的績效，並透過思想領導力和信譽良好的合作夥伴關係建立起知識權威，兩者相結合，能夠建立強大且必要的信任機制。這種雙重方法既滿足了對已證實有效性的直接營運需求，也滿足了對可靠、合乎道德且尖端專業知識的長期戰略需求，從而有效降低了未經驗證的人工智慧解決方案帶來的風險。

**5.**策略訊息傳遞和品牌協調

為「數據探索者」精心設計的策略訊息旨在與情報機構合作時發揮最大影響力和適用性，不斷強化其核心價值主張和使命。

所提出的口號和標語特別有效：

- "數據探索者：穿越人工智慧雪崩，獲取真實情報。"這句口號非常有力，直接傳達了核心問題——「人工智慧雪崩」——以及提出的解決方案：「導航……實現真正的智慧」。[1]「數據探索者」一詞本身獨特而令人難忘，讓人聯想到深入探索複雜數據領域並充滿挑戰的景象。「真實情報」則直擊了假訊息氾濫時代對真實訊息的迫切需求。這句口號簡潔明了，直擊了目標受眾在人工智慧驅動的世界中對數據真實性的擔憂，因此非常貼切。[1]
- "在數字洪流中發掘真相：道德情報的數據探索者。"「挖掘真相」是一個強而有力的短語，與情報機構的基本使命產生了深刻的共鳴。[1]「數位洪流」是另一個對海量資料的有效隱喻。「情報倫理」這個關鍵補充對於建立信任以及適應現代情報界不斷發展的價值觀和要求至關重要。[1]這句口號非常貼切，既強調了發現真相，也強調了

道德層面，這也是「資料探索者」的關鍵區別因素。[1]

- “為全球和平提供真實情報：數據探索者的承諾。”這句標語直接將該服務與「全球和平」這一最終的高層次目標聯繫起來，這是道德情報機構的主要目標。[1]「真實情報」強化了核心價值主張，而「資料探索者的承諾」則增添了一層承諾和可信度。雖然這個口號恰如其分，但它更側重於結果和品牌承諾，而非「數據探索者」的直接行動。[1]

整體而言，這些口號有力、有影響力，並且非常貼合目標受眾。它們有效地運用了引人入勝的比喻來描述問題，並強調了「數據探索者」在提供可驗證且合乎倫理的情報方面的獨特作用。[1]

關鍵資訊點同樣經過精心設計，旨在與情報機構的優先事項產生共鳴：

- “即將到來的數據雪崩威脅著全球情報的完整性。”這個資訊點具有極強的影響力，立即凸顯了問題的迫切性和嚴重性。它直指情報機構的核心職能和關切，使其成為從國家安全和全球穩定的角度闡述問題、設定議題的絕佳聲明。[1]

- “只有專業的‘數據探索者’才能深入挖掘並驗證關鍵信息。”這句話具有很強的衝擊力，將「數據洞穴探險者」定位為唯一且必要的解決方案。[1] 「只專注於專業」這個短語營造出一種獨特的專業感，而「深入挖掘以驗證」則清晰地定義了他們的核心能力。強調「關鍵訊息」則凸顯了所涉及的高風險。這個資訊點非常貼切，因為它清楚地表達了獨特的價值主張，並將「資料探索者」與更一般的資料分析服務區分開來。[1]
- “我們為道德情報機構提供可驗證的數據，以保障和平與安全。”這個訊息點非常強烈，專注於客戶的直接利益（「授權」），並明確地將他們的服務與「和平與安全」的總體目標聯繫起來。[1]「合乎道德」和「可驗證資料」的加入強化了他們的核心原則和可信度。這是一個絕佳的宣傳點，因為它將「資料探索者」的使命與情報機構的更廣泛目標一致，使他們的服務顯得對於實現這些目標不可或缺。[1]
- “我們對道德人工智慧的承諾確保了信任和負責任的決策。”這個資訊點影響很大，直接解決了人工智慧時代的一個關鍵問題：道德和信任。[1] 它向潛在客戶保證，「數據探索者」的經營誠信，

> 其方法能夠帶來可靠的結果。這非常合適，因為在人工智慧可能被用於惡意目的的環境中，強調道德原則是一個強大的差異化因素，並能建立信譽。[1]

關鍵訊息點強有力、影響力十足，且高度契合目標受眾。它們有效地界定了問題，定位了解決方案，突出了優勢，並強化了核心價值觀，同時使用了符合情報機構優先事項的語言。[1]在整個資訊傳遞過程中始終強調「真實性」、「驗證」和「道德」是一大優勢，直接解決了人工智慧驅動的虛假資訊所帶來的核心挑戰，並建立了基礎信任。[1]這項策略訊息超越了單純的技術服務。「資料探索者」將此問題視為全球情報的生存威脅，並將自己定位為維護和平與安全不可或缺且合乎道德的解決方案，他們並非僅僅在銷售工具，而是在倡導一種至關重要的伙伴關係。這種方法對於確保高層機構的參與和投資非常有效，因為它將他們的價值主張與情報界的根本目標和生存關切直接聯繫起來。

打擊人工智慧驅動的假訊息充滿了重大且不斷變化的挑戰，影響著技術能力、威脅的本質以及道德法律框架。

動態數位生態系中的技術挑戰：
在數位生態系統中，從社群媒體貼文到多媒體內容，每天都會產生大量且快速的內容，這構成了巨大的障礙。3 假訊息以驚人的速度傳播，往往超出了傳統事實查核的力度。病毒式傳播的假訊息可以在數小時內傳播到數百萬人，而即使更正資訊已經發布，也難以達到類似的滲透率。3 這種不平衡現象迫切需要開發和部署可擴展的自動化工具，以便即時處理和驗證大量資料。3

目前虛假資訊檢測工作的關鍵限制是缺乏多樣化和動態的資料集。現有資料集通常不足，迫切需要特定平台和特定語言的資料。不同文化、平台和模式下的細微差別和脈絡尚未充分體現。[3]大多數現有資料集主要側重於文本，在多模態檢測能力方面存在顯著差距。此外，新的主題和虛假資訊形式不斷湧現，尤其是在疫情、選舉或衝突等全球性事件期間，這要求資料集能夠動態且持續更新。[3]對 X、Instagram 和 Facebook 等主要社群媒體平台的數據存取也經常受到限制，進一步阻礙了全面分析。[3]多模態假訊息活動日益盛行，它們巧妙地融合了文字、圖像和視頻，以增強可信度和參與度，這又增加了一層複雜性。偵測和分析這種多模態敘事需要複雜的跨模態人工智慧系統，能夠關聯不同格式的訊息，這是一項既複雜又耗費資源的任務。[3]

不斷演變的人工智慧驅動威脅：

生成式人工智慧的最新進展，尤其是大型語言模型和生成對抗網路 (GAN)，使得創建極具說服力的虛假內容成為可能，從而大大加劇了這一挑戰。3 這種複雜性使人類分析師和現有的自動化工具越來越難以辨別真實性。3 除了簡單的內容生成之外，高階演算法現在還能夠進行超定向宣傳，分析用戶資料以精確傳遞虛假訊息，利用個人偏見或弱點。5 此外，演算法本身存在漏洞，有偏見或不透明的人工智慧系統可能會被有意或無意地用來壓制合法內容或放大有害的錯誤訊息。5

道德與法律挑戰：

人工智慧在假資訊檢測中的應用引發了嚴重的倫理問題，尤其是在資料隱私和演算法偏見方面。6 使用有偏見的資料訓練的人工智慧模型可能會導致歧視性結果，自動評估系統不成比例地懲罰來自低收入社區的學生，或者課程推薦系統表現出性別偏見的案例就證明了這一點。6 為防止意外的審查，必須對內容審核演算法進行嚴格的偏見測試。5 解決人工智慧驅動的虛假資訊的法律機制的未來在很大程度上仍未確定，這給立法者帶來了複雜性，他們必須協調國家法律框架與監管潛在危險數位內容的必要性。7 不斷發展的人工智慧政策和監管格局為機構帶來了不確定性，可能導致人工智慧計畫的延遲或變化。8 此外，許多複雜的人工智慧模型缺乏透明度，使得它們的決策過程難以理解。這種不透明性引發了人們對問責和信任的嚴重擔憂。9 因此，強大的文件和可追溯性對於審計和治理至關重要。5 保護人工智慧訓練資料和模型存取對於防止篡改或濫用也至關重要，因此需要採取加密計算和聯邦學習等措施。5

公眾信任的侵蝕：

隨著假訊息變得越來越複雜和普遍，它系統性地侵蝕了公眾對媒體、政府機構以及民主話語基礎的信任。5 這種信任的侵蝕代表著一項重大的社會責任，而負責任的人工智慧發展旨在應對這項責任。5

下表以結構化、高層次的方式概述了情報機構和「資料探索者」等專業機構必須面對的多方面挑戰。表格將這些挑戰分為技術層面、不斷演變的威脅以及道德/法律層面，以便快速理解作戰環境的複雜性。對於決策者而言，它凸顯了問題的系統性，並強調了為何需要全面且專業的解決方案。

## 表 2：打擊人工智慧假資訊的主要挑戰

| 挑戰類別 | 具體挑戰 | 意義/影響 |
| --- | --- | --- |
| 技術限制 | 內容的數量和速度、缺乏多樣化/ | 超越事實查核，需要可擴展的工具； |

| | | |
|---|---|---|
| | 動態資料集、多模式虛假訊息 | 限制適用性、實用性和即時適應性；需要複雜的跨模式人工智慧系統。 |
| 不斷演變的人工智慧威脅 | 產生人工智慧的複雜性、超針對性宣傳、演算法漏洞 | 難以辨別真實性；利用個人偏見；可以壓制合法內容或放大錯誤訊息。 |
| 道德與法律問題 | 資料隱私與演算法偏見、監管不確定性、透明度、問責制、安全性 | 導致歧視性結果、意外的審查；延遲人工智慧計劃，產生合規風險；削弱信任，阻礙審計；存在篡改/濫用的風險。 |
| 社會影響 | 公眾信任的侵蝕 | 破壞媒體、政府機構和民主話語；這是一項重大的社會需求。 |

這裡描述的動態體現了一場固有的「人工智慧軍備競賽」。隨著人工智慧驅動的虛假資訊日益複雜且適應性強，檢測方法必須不斷發展並預測新的威脅，

而不僅僅是對現有威脅做出反應。這意味著任何有效的解決方案，包括「數據探索者」提供的解決方案，不僅必須具備現有能力，還必須展現出強大的持續研究、開發和主動適應能力，以保持領先地位。這代表著一項長期的策略承諾，而非一次性的技術解決方案。

**7.** 情報機構採用人工智慧的實施障礙

除了人工智慧假訊息帶來的外部挑戰之外，情報機構在採用和有效利用人工智慧解決方案方面還面臨著許多內部障礙。這些組織和基礎設施的考量對於成功實施至關重要。

解決人工智慧人才缺口：聯邦政府實現人工智慧準備面臨的最大挑戰之一是難以找到並留住熟練的人工智慧人才。[8]由於與私營部門的激烈競爭，各機構難以吸引熟練的人工智慧專業人員。[8]專業知識的缺乏直接限制了機構有效設計、實施和管理人工智慧計畫的能力。[8]要克服勞動力缺口，就必須對現有員工進行策略性投資培訓，並創造引人注目的激勵措

施來吸引頂尖人工智慧人才加入公共服務。[8]

確保資料品質和安全：任何人工智慧解決方案的成功實施從根本上都依賴高品質、管理良好的數據。[8]不幸的是，許多機構所處理的數據不完整、不準確或不一致。[9]資料治理不善會直接導致人工智慧輸出不準確，尤其是在時間敏感的操作中，這會嚴重降低人們對人工智慧驅動決策的信任。[8]資料孤島、不同的資料格式以及對遺留系統的依賴進一步為人工智慧應用存取和有效利用資料設置了重大障礙。[9]解決方案包括建立強大的資料治理框架、提高跨系統的互通性以及投資資料整合平台。[8]此外，保護人工智慧訓練資料和模型存取對於防止篡改或濫用至關重要，需要安全的運算環境。[5]

探索人工智慧監管環境：人工智慧政策和法規的動態制定和演變對聯邦政府內部人工智慧的採用提出了嚴峻挑戰。[8]行政命令和監管框架的變化可能會導致人工智慧計畫的延遲或重大變化，造成不確定的環境。[7]各機構必須在其人工智慧策略中培養敏捷性和適應性，以確保持續遵守新興法規。[8]

成本影響與投資報酬率衡量：實施人工智慧解決方

案的成本可能很高，需要仔細評估初始成本和潛在投資回報率 (ROI)。[9]高昂的初始成本和持續的開支使得採取策略方法成為必要，例如從規模較小、可控的試點計畫開始，以證明其價值並確保獲得更大規模計畫的支持。[9]探索基於雲端的人工智慧解決方案有助於降低基礎架構成本，而優化雲端資源利用率對於管理持續開支至關重要。[9]從一開始就為人工智慧專案定義明確的目標和指標對於有效追蹤其對關鍵績效指標的影響並向利害關係人傳達結果至關重要。[9]

與現有 IT 基礎架構和工作流程中斷的兼容性：人工智慧系統可能與現有的 IT 基礎設施或遺留系統不相容，通常需要進行重大修改或開發客製化整合解決方案。[9]人工智慧的引入也會擾亂既定的工作流程和流程，需要謹慎的變革管理和全面的員工培訓，以確保平穩過渡並培養採用人工智慧的文化。[9]

下表綜合展現了情報機構在採用人工智慧解決方案時所面臨的內部、組織和基礎設施挑戰。這有助於決策者理解，問題不僅在於找到合適的外部解決方案，還在於內部的準備情況。對於“數據探索者”，該表格重點介紹了他們可能需要提供補充服務或策略

建議以促進採用的領域。

表 3：情報機構實施人工智慧的障礙

| 跨欄項目 | 具體挑戰 | 對人工智慧採用的影響 |
| --- | --- | --- |
| 勞動力和專業知識 | 人工智慧人才缺口 | 限制有效設計、實施和管理人工智慧計畫的能力。 |
| 數據基礎設施 | 資料品質與安全性差、資料孤島、遺留系統 | 導致人工智慧輸出不準確，降低信任度；為有效利用數據設定障礙。 |
| 治理與政策 | 人工智慧監理旋風 | 造成人工智慧計畫的不確定性、潛在的延遲或變化。 |
| 財務與價值 | 初始成本高，持續費用高，投資報酬率難以衡量 | 需要仔細的成本效益分析；阻礙獲得更大措施的支持。 |
| 整合與營運 | 相容性問題、工作流程中斷 | 需要進行重大修改/客製化解決方案；需要謹慎的變 |

| | | |
|---|---|---|
| | | 更管理和員工培訓。 |

即使是最先進的外部人工智慧解決方案（例如「資料探索者」提供的方案），其有效性也在很大程度上取決於客戶機構內部組織的準備程度。缺乏熟練的人才、數據品質受損或監管環境不穩定，都可能嚴重阻礙任何人工智慧驅動的智慧的實用性和可信度。這意味著「資料探索者」不僅要提供尖端服務，還要準備好為更廣泛的機構工作提供建議或參與其中，以建立基礎人工智慧準備，從而使雙方的合作成為更全面的合作關係，而非簡單的供應商-客戶關係。

**8.** 市場格局與競爭定位

人工智慧驅動的假訊息檢測解決方案市場具有獨特的雙重性，並且對可信度有著迫切的需求。人工智慧在這一領域扮演著雙重角色：它既是製造複雜虛假內容的強大工具，也是檢測和打擊虛假內容的不可或缺的手段。[2]這種動態促使研究人員、科技公司和政府之間的合作，共同利用人工智慧技術打擊人

工智慧驅動的錯誤訊息。[2]

在這種情況下，我們迫切需要專業化、可擴展的解決方案。數位內容的大量和快速成長速度需要能夠即時處理和驗證大量資料的自動化工具。[3]這凸顯了市場對能夠有效管理「資料雪崩」並提供可操作情報的高度專業化解決方案的巨大需求。

然而，人工智慧服務市場並非沒有陷阱，凸顯了可信度的重要性。聯邦貿易委員會（FTC）等監管機構已對發布虛假人工智慧聲明的公司採取了行動。[10]例如“DoNotPay”虛假宣傳自己是“世界上第一個機器人律師”，但沒有提供任何有效性證據；“Ascend Ecom”則誤導性地宣稱其利用“尖端”人工智慧工具來產生被動收入。[10]這些案例強調了人工智慧解決方案市場中證據支持的主張、道德行為和可驗證結果的至關重要性。

儘管面臨諸多挑戰，目前的環境也為合作與夥伴關係提供了重要的機會。打擊人工智慧驅動的虛假資訊本質上是一項集體努力，平台公司需要與專業的事實查核人員和內容審核員合作，社會科學家則需要積極研究虛假資訊。[2]雅典娜計畫等倡議強調了將

人工智慧創新與強有力的保障措施相結合的迫切需要，倡導建立以公平、透明、問責和安全原則為基礎的框架。[5]這種合作精神創造了有利於建立策略夥伴關係的環境，例如「資料洞穴探險者」與學術機構和道德人工智慧智庫提出的夥伴關係。[1]在競爭激烈且可能存在欺騙性的人工智慧解決方案市場中，「資料探索者」對合乎道德的人工智慧和可驗證身分驗證的堅定承諾，不僅是一種道德立場，更是一種關鍵的競爭優勢。對於在高風險環境中運作的情報機構來說，這無疑是一種強大的信任訊號，因為準確性、可靠性和誠信至關重要。這種道德差異化使他們成為值得信賴且負責任的合作夥伴，從而有別於那些缺乏道德操守或未經驗證的人工智慧供應商。

## 9. 建議與策略展望

為了進一步增強其產品和市場滲透力，「數據探索者」應該考慮以下建議：

為「數據探索者」提供增強其產品和市場滲透力的建議：

- 確定優先次序並擴大試點計畫：鑑於試點計畫對建立信任和展示能力的巨大影響，「數據探索者」應該積極與情報機構合作，並宣傳成功的試點計畫。[1]這些專案應該經過精心設計，直接解決機構最迫切的現實數據挑戰，提供實際的價值證明。
- 深化人工智慧道德領導：繼續發佈白皮書，為行業標準做出貢獻，並積極參與以人工智慧倫理為重點的高調會議。[1]探索開源道德人工智慧工具或框架的開發可以進一步鞏固其領導地位並展示透明度，促進更廣泛的信任和合作。
- 解決機構實施障礙：雖然他們的核心服務是外部的，「數據洞穴探險者」應該考慮提供諮詢服務或建立合作夥伴關係，並專注於幫助機構克服內部人工智慧採用障礙。[8]這可能包括提供資料治理的最佳實踐、吸引和留住人工智慧人才的策略，或提供專門的培訓和共同開發計劃。
- 主動威脅預測：認識到「人工智慧軍備競賽」的動態，持續投資研發對於預測人工智慧產生的虛假資訊的新形式至關重要。[3]這種積極主動的方法確保他們的能力始終領先於不斷演變的威脅，並保持競爭優勢。

- 擴展多式聯運能力：明確強調並進一步投資於偵測和分析多模式虛假資訊的先進能力。[3]隨著虛假宣傳活動越來越多地融合文字、圖像和視頻，強大的跨模態分析將是一個日益嚴峻且複雜的挑戰，需要專業知識。

情報機構採用此類服務的策略考量：

- 優先考慮道德框架：各機構必須堅持要求任何人工智慧解決方案提供者遵守明確的人工智慧道德原則、透明度和責任制。[1]這不僅僅是一個合規問題，也是維護公眾信任、確保情報可靠以及減輕偏見或濫用風險的基本要求。
- 投資基礎資料準備：在部署先進的人工智慧解決方案之前，機構必須優先提高內部資料品質、建立強大的資料治理框架以及增強跨系統的互通性。[8]數據品質差或不一致必然會削弱即使是最複雜的人工智慧的有效性。
- 培養人工智慧素養和人才發展：積極投資培訓現有人員並創造激勵措施來吸引頂尖人工智慧人才至關重要。[8]熟練的內部員工對於有效整合、管理和細緻解讀人工智慧驅動的智慧至關重要。

- 推行盡職調查試點計畫：各機構應利用試點計畫作為評估人工智慧解決方案的主要機制。[1]這樣就可以在受控環境中對能力進行直接、實際的評估，在全面採用之前提供可驗證的有效性證據。
- 尋求策略夥伴關係：積極尋求與「資料洞穴探險者」等專業實體的合作可以顯著增強內部能力，特別是在人工智慧驅動的虛假資訊檢測等利基和快速發展的領域。[1]

人工智慧驅動的資訊環境下智慧未來的長期展望：
「數據雪崩」和人工智慧生成內容的激增代表著全球資訊格局的永久性和根本性轉變。情報的未來將越來越取決於駕馭這種極其複雜的環境、嚴格驗證資訊真實性以及堅定維護公眾信任的能力。像「經驗證的倫理情報」這樣的解決方案並非權宜之計，而是建構一個強大、有韌性、符合倫理道德的情報體係不可或缺的持久組成部分。這對於長期維護全球穩定與和平至關重要。人工智慧威脅的不斷演變要求我們始終致力於制定適應性策略，促進跨學科合作，並堅持堅定不移的倫理原則，以確保資訊的完整性。

## 10. 結論

人工智慧產生的虛假訊息即將引發“數據雪崩”，對全球情報的完整性和民主話語的基礎構成了前所未有的生存威脅。不可否認，人們迫切需要掌握和驗證數位資訊的專業知識，因此「經認證的道德情報」在這一不斷變化的情況下成為不可或缺的應對措施。

「數據探索者」團隊提出了一個令人信服且清晰的價值主張。他們將高級資料取證和真實性驗證的深厚技術專長與對人工智慧倫理原則的堅定而明確的承諾相結合。他們的戰略訊息致力於為各機構提供可驗證的數據，以促進和平與安全，這與情報界的最高優先事項直接契合。儘管在打擊人工智慧虛假資訊以及在政府機構內部實施人工智慧解決方案方面依然面臨重大挑戰，「數據探索者」團隊提出的方法論、全面的信任建構策略以及道德差異化策略，使他們成為重要且值得信賴的合作夥伴。他們的服務不僅僅是識別虛假資訊；他們的根本目標是維護知情決策、預防衝突以及持久追求全球穩定與和平所必需的真相基礎。

參考文獻


1. 人工智慧為公共互聯網提供層層自我資訊...
2. 人工智慧與虛假資訊 - 2024 年院長報告，瀏覽日期：2025 年 7 月 6 日 https://2024.jou.ufl.edu/page/ai-and-misinformation
3. 假資訊偵測中的人工智慧 - 應用網路安全與網路治理，存取日期：2025 年 7 月 6 日 https://www.acigjournal.com/AI-in-Disinformation-Detection,200200,0,2.html
4. 產生人工智慧的興起與網路虛假新聞和虛假資訊的威脅：性醫學的觀點，造訪日期：2025 年 7 月 6 日，https://pmc.ncbi.nlm.nih.gov/articles/PMC11076802/
5. 使用負責任的人工智慧打擊假訊息 - Trilateral Research，造訪日期：2025 年 7 月 6 日 https://trilateralresearch.com/responsible-ai/using-responsible-ai-to-combat-misinformation
6. 生成式人工智慧在教育領域的倫理與監管挑戰：系統性綜述，造訪日期：2025 年 7 月 6 日https://www.frontiersin.org/articles/10.3389/feduc.2025.1565938
7. 透過國家監管打擊人工智慧驅動的虛假訊息：從烏克蘭案例中學習 - Frontiers，造訪日期：2025 年 7 月 6 日，https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.202

4.1474034/full
8. 聯邦政府採用 AI/ML 需要克服的 3 個挑戰，造訪日期：2025 年 7 月 6 日，https://fedtechmagazine.com/article/2025/07/3-challenges-overcome-aiml-adoption-federal-government
9. 《人工智慧實施的障礙：企業因應挑戰》，瀏覽日期：2025 年 7 月 6 日，https://www.ml-science.com/blog/2025/2/26/the-hurdles-of-ai-implementation-navigating-the-challenges-for-enterprises
10. FTC 宣布嚴厲打擊欺騙性人工智慧索賠和陰謀，造訪日期：2025 年 7 月 6 日，https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes